Chapter 1
A Discovery Method of Attractive Rules from the Tabular Structured Data

Shigeaki Sakurai
Tokyo Institute of Technology, Japan

ABSTRACT
This chapter introduces a discovery method of attractive rules from the tabular structured data. The data is a set of examples composed of attributes and their attribute values. The method is included in the research field discovering frequent patterns from transactions composed of items. Here, the transaction and the item are a receipt and a sales item in the case of the retail business. The method focuses on relationships between the attributes and the attribute values in order to efficiently discover patterns based on their frequencies from the tabular structured data. Also, the method needs to deal with missing values. This is because parts of attribute values are missing due to the problems of data collection and data storage. Thus, this chapter introduces a method dealing with the missing values. The method defines two evaluation criteria related to the patterns and introduces a method that discovers the patterns based on the two-stepwise evaluation method. In addition, this chapter introduces evaluation criteria of the attractive rules in order to discover the rules from the patterns.

DOI: 10.4018/978-1-4666-1806-0.ch001
INTRODUCTION

Owing to the progress of the computer environment and the network environment, it is easier and easier to collect and store large amounts of data. In the near future, we believe that many sensors are buried in real world environments and are attached to ourselves. Also, we believe that they compose huge sensor networks. The data related to our daily life will be easily collected and stored through the networks. We can anticipate that the analysis of the data leads to revise our daily life and to realize more smart society. Therefore, various types of the analysis have been aggressively studied since mid-1990s. The research field expands more and more. This is because there are various types of the data and there are various types of analysis needs. However, the analysis depends on both the data types and the needs to some extent. It is impossible to construct only one analysis method. This chapter focuses on the discovery of frequent patterns and association rules from the collected data as one of the analysis tasks. The patterns are frequently observed in the data. The rules are discovered from the patterns and represent characteristic combinations of items. Here, an item is a minimum element composing the data. The analysts can recognize relationships hidden in the data by checking the patterns and the rules. They can activate the patterns and the rules to their decision making. For example, in the case of the retail businesses, sales managers can use the patterns and the rules to decide how to display sales items. Also, in the case of the information security field, the system administrators can use the patterns and the rules to analyze attack methods of the crackers. On the other hand, the discovery of the patterns and the rules is a difficult task because we have to discover them from exponential combinations of items and to process large amounts of data. It is necessary to efficiently discover the patterns and the rules.

Thus, in the following, this chapter firstly introduces related works in this research field. It shortly introduces two representative discovery methods of frequent patterns. Next, it focuses on the discovery method of patterns from the tabular structured data. The data is a set of examples composed of attributes and their attribute values. The method focuses on relationships between the attributes and the attribute values in order to efficiently discover the patterns. Also, this method focuses on missing values in order to discover more valid patterns. In addition, this chapter introduces evaluation criteria in order to extract attractive rules from the patterns. Lastly, this chapter introduces the future trend of this research field.

BACKGROUND

Basket analysis of receipts collected from the retail business is the origin of the discovery of both frequent patterns and association rules. The rules are usually extracted from the discovered patterns. Therefore, it is important to efficiently discover the patterns. Each receipt is defined as a transaction in the analysis. Then, each transaction is composed of some items such sales items in the retail business. Each item is regarded as either of two cases in the transaction. That is, one case shows that the item is included in the transaction and the other case shows that the item is not included. Agrawal and Srikant (1994) and Han et al. (2000) propose representative discovery methods of frequent patterns using the monotonic property of the patterns. The property shows that if the pattern grows, its evaluation criterion monotonically decreases. It is called the Apriori property. On the other hand, Morzy and Zakrzewicz (1998) and Zaki et al. (1997) propose methods that speedily discover the patterns by devising storage methods for the data. Also, Koh et al. (2005) proposes a method that discovers association rules with low support but high confidence. Here, the support and the confidence are evaluation criteria of the patterns and the rules. Yan et al. (2005) proposes